Class 06: Model Evaluation & Model Selection

Martin Slawski

Volgenau School of Engineering
Department of Statistics

March 8th, 2018

Model Evaluation & Model Selection

Model Evaluation:

Given some fixed function f to be used for prediction in regression or classification, we would like to assess the prediction performance of f, with respect to a certain **performance measure**.

Model Evaluation & Model Selection

Model Selection:

Given a set $\{f_\mu\}_{\mu \in \mathcal{M}}$ of such functions ("models"), pick the one that does best.

Model Averaging:

Combine $\{f_\mu\}_{\mu \in \mathcal{M}}$ in a such way that the aggregated results improves over the best one could get by using a single function.

In statistics, this is also known as **aggregation**.

Evaluation of binary classifiers

Given labels $Y_1, \ldots, Y_n$, and predictions $\widehat{Y}_1, \ldots, \widehat{Y}_n$ we can compute the so-called **confusion matrix**:

	$\widehat{Y} = 1$	$\widehat{Y} = -1$	
$Y = 1$	TP	FN	P
$Y = -1$	FP	TN	N

True positive rate (or **recall** or **sensitivity**):

$$\text{TPR} = \frac{\text{TP}}{P}$$

False positive rate:

$$\text{FPR} = \frac{\text{FP}}{N}$$

True negative rate (or **specificity**):

$$\text{TNR} = \frac{\text{TN}}{N}$$

False negative rate:

$$\text{FNR} = \frac{\text{FN}}{P}$$

Evaluation of binary classifiers

Positive predictive value (or **precision**):

$$\text{PPV} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

Negative predictive value:

$$\text{NPV} = \frac{\text{TN}}{\text{TN} + \text{FN}}$$

The standard "summary measure" is the (misclassification) error rate:

$$\text{err} = \frac{\text{FP} + \text{FN}}{n}$$

Evaluation of binary classifiers

The misclassification rate is not a good performance measure if the class probabilities are highly unbalanced.

This setting is in fact not too rare as indicated by the following examples:

- anomaly detection,
- rare diseases,
- problems in information retrieval:

  find text documents or images related to a specific topic among a large collection

If one of the two classes is strongly underrepresented it is trivial to achieve low error. In this case, it is bad practice to report only the error rate.

Evaluation of binary classifiers

A good classifier achieves an appropriate balance between the TPR and the FPR, or equivalently, between the TPR and the PPV (or, synonymously **recall** and **precision**).

In this context, a standard summary measure is the F-score

$$\mathsf{F} = 2\frac{\mathsf{PPV} \cdot \mathsf{TPR}}{\mathsf{PPV} + \mathsf{TPR}},$$

the harmonic mean between precision and recall.

ROC curves

Another interesting concept is the ROC (**R**eceiver **O**perator **C**haracteristic) curve for *real-valued* f.

With f, we can define the following set of classifiers:

$$g_t(x) = \text{sign}(f(x) - t), \quad t \in \mathbb{R},$$

and then evaluate how FPR and TPR change with t.

ROC analysis is widely used in Biostatistics for evaluating and comparing biomarkers.

ROC curves

Given a (test) sample $\{(X_1, Y_1), \ldots, (X_n, Y_n)\}$, $Y_i \in \{-1, 1\}$, $i = 1, \ldots, n$, let us use the shortcut

$$f_i = f(X_i), \;\; i = 1, \ldots, n.$$

Let $\mathrm{FPR}(t)$ and $\mathrm{TPR}(t)$ denote the true respectively false positive rate associated with the predictions

$$\widehat{Y}_i = \widehat{Y}_i(t) = \mathrm{sign}(f_i - t),$$

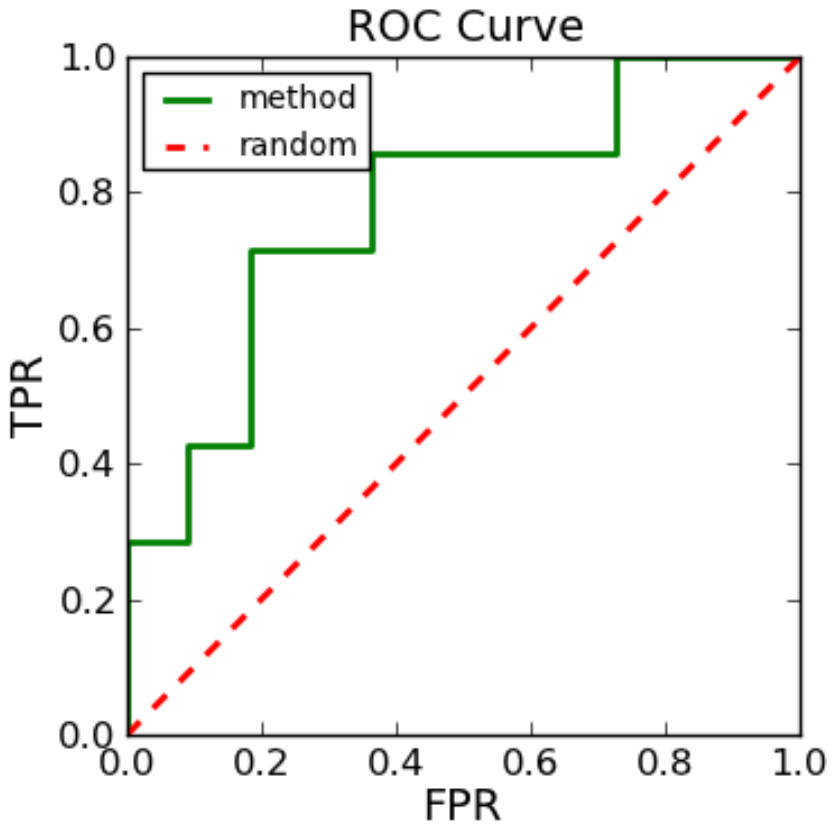
with the convention that $\mathrm{sign}(0) := 1$.

The ROC is then defined as the piecewise constant function passing through the points

$$\{(\mathrm{FPR}(t), \mathrm{TPR}(t))\}_{t \in \{f_1, \ldots, f_n, \infty\}}.$$

ROC curves

The ROC is then defined as the piecewise constant function passing through the points $\{(\text{FPR}(t), \text{TPR}(t))\}_{t \in \{f_1, \ldots, f_n, \infty\}}$.



ROC curves

Properties of the ROC curve:

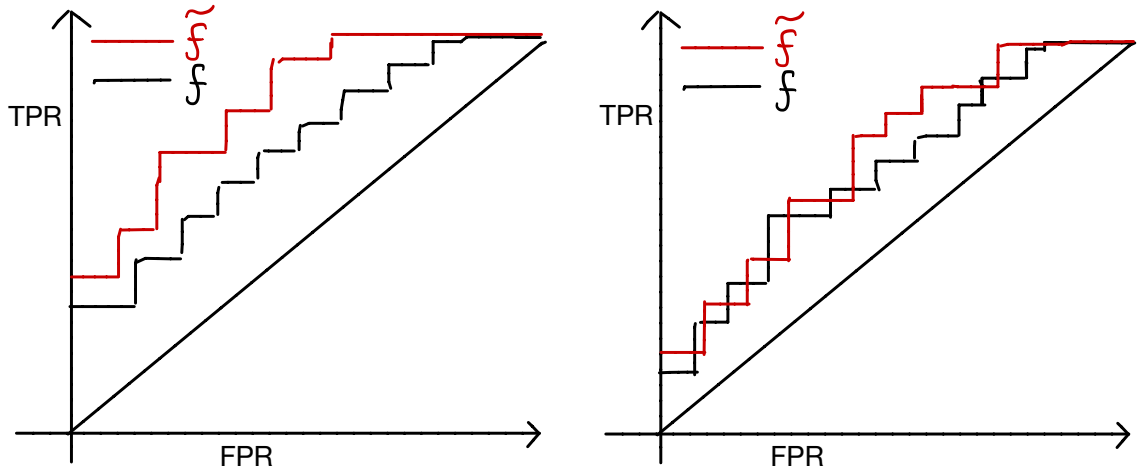
- it passes through the points $(0, 0)$ and $(1, 1)$
- it is monotonically increasing
- it is invariant under monotonically increasing transformations applied to $\{f_i\}_{i=1}^n$
- the ROC of a "perfect classifier" passes through $(0, 1)$
- the ROC of the random classifier

$$\begin{cases} 1 & \text{with probability } \mathbf{P}(Y = 1), \\ -1 & \text{with probability } \mathbf{P}(Y = -1), \end{cases}$$

corresponds to the angle bisector.

ROC curves

Comparison of two classifiers f, $\widetilde{f}$ based on their ROC curves:



In the left plot, $\widetilde{f}$ is "uniformly better" than f.

This is not the case for the right plot. Nevertheless, f and $\widetilde{f}$ can still be compared in terms of the **AUC** (Area under the curve).

- the higher the AUC, the better,
- a "perfect classifier" achieves an AUC of 1,
- a random classifier achieves an AUC of 0.5.

Statistical Inference for the Generalization Error

We want to estimate

$$\text{err} = \mathbf{E}_{X,Y}[I(Y \neq \widehat{f}(X))|\mathcal{D}_n],$$

where $\widehat{f}$ is a classifier learned based on $\mathcal{D}_n = \{(X_i, Y_i)\}_{i=1}^n$.

We have seen that empirical error (with respect to $\mathcal{D}_n$)

$$\text{err}_{\mathcal{D}_n} = \frac{1}{n}\sum_{i=1}^{n} I(Y_i \neq \widehat{f}(X_i))$$

underestimates err, often dramatically.

Statistical Inference for the Generalization Error

Suppose we are given a data set $\mathcal{D}'_m = \{(X'_i, Y'_i)\}_{i=1}^m$ of size m drawn from the same distribution as $\mathcal{D}_n$.

Then the random variables $\{I(Y'_i \neq \widehat{f}(X'_i))\}_{i=1}^m$ are i.i.d. Bernoulli random variables with probability of success given by

$$\text{err} = \mathbf{E}_{X,Y}[I(Y \neq \widehat{f}(X))|\mathcal{D}_n].$$

This suggests the estimator

$$\text{err}_{\text{hold-out}} = \frac{1}{m}\sum_{i=1}^m I(Y'_i \neq \widehat{f}(X'_i)).$$

Statistical Inference for the Generalization Error

We can bound the deviation between and $\text{err}_{\text{hold-out}}$ and err using Hoeffding's inequality.

Hoeffding's inequality:

Let $\{Z_i\}_{i=1}^m$ be i.i.d. random variables contained in some interval $[a, b]$ with probability one, and let $\bar{S}_m = \frac{1}{m}\sum_{i=1}^m Z_i$. Then for all $\varepsilon > 0$

$$\mathbf{P}\left(\left|\bar{S}_m - \mathbf{E}[\bar{S}_m]\right| \geq \varepsilon\right) \leq 2\exp\left(-2m\varepsilon^2/(b-a)^2\right).$$

We will use this result with

$$Z_i = I(Y_i' \neq \widehat{f}(X_i')), \quad i = 1, \ldots, m,$$

so that

$$\bar{S}_m = \text{err}_{\text{hold-out}}, \qquad \mathbf{E}[\bar{S}_m] = \text{err}.$$

Statistical Inference for the Generalization Error

We hence obtain that with probability at least $1 - \delta$

$$|\text{err}_{\text{hold-out}} - \text{err}| \leq \sqrt{\frac{\log(2/\delta)}{2m}}$$

Alternatively, in order to have

$$|\text{err}_{\text{hold-out}} - \text{err}| \leq \varepsilon,$$

we need $m \geq \frac{1}{2\varepsilon^2} \log\left(\frac{2}{\delta}\right)$.

Statistical Inference for the Generalization Error

The bound on $|\text{err}_{\text{hold-out}} - \text{err}|$ is straightforward to compute, but it is possible to obtain a sharper result, using the **Clopper-Pearson** confidence interval for proportions:

For $p \in [0,1]$, let $F_{m,p}$ denote the CDF of a $\text{Bin}(m,p)$-random variables (m:#trials, p: probability of success), and consider

$$\widehat{p} = \frac{1}{m} S_m, \quad \text{where } S_m \sim \text{Bin}(m, p^*).$$

Define $\underline{p}$ and $\overline{p}$ by the relations

$$F_{m,\underline{p}}(m\widehat{p}) = 1 - \frac{\delta}{2}, \qquad F_{m,\overline{p}}(m\widehat{p}) = \frac{\delta}{2}.$$

Then $[\underline{p}, \overline{p}]$ is an $1-\delta$ confidence interval for p^*, i.e,

$$\mathbf{P}_{p^*}(\underline{p} \le p^* \le \overline{p}) \ge 1 - \delta.$$

Statistical Inference for the Generalization Error

Comparison of two classifiers based on a statistical test:

Suppose we are given two classifiers $\widehat{f}$, $\widetilde{f}$, and based on their hold-out errors, we want to test whether their generalization errors

$$\text{err} = \mathbf{E}_{X,Y}[I(Y \neq \widehat{f}(X))|\mathcal{D}_n], \qquad \widetilde{\text{err}} = \mathbf{E}_{X,Y}[I(Y \neq \widetilde{f}(X))|\mathcal{D}_n]$$

are different, i.e., we consider the test hypotheses

$$H_0 : \text{ err} = \widetilde{\text{err}}, \qquad H_1 : \text{ err} \neq \widetilde{\text{err}}.$$

Statistical Inference for the Generalization Error

Test I: asymptotic t-test

Let denote

- $Y'_1, \ldots, Y'_m$ the labels in the test set,
- $\widehat{Y}_1, \ldots, \widehat{Y}_m$ the predictions made my classifiers $\widehat{f}$,
- $\widetilde{Y}_1, \ldots, \widetilde{Y}_m$ the predictions made my classifiers $\widetilde{f}$.

Define $W_i = I(\widehat{Y}_i \neq Y'_i) - I(\widetilde{Y}_i \neq Y'_i)$, $i = 1, \ldots, m$. We have

$$\frac{1}{m} \sum_{i=1}^{m} W_i = \text{err}_{\text{hold-out}} - \widetilde{\text{err}}_{\text{hold-out}},$$

where $\widetilde{\text{err}}_{\text{hold-out}}$ denotes the hold-out error of $\widetilde{f}$.

Statistical Inference for the Generalization Error

Test I: asymptotic t-test

We then use the test statistic

$$T = \sqrt{m}\frac{\frac{1}{m}\sum_{i=1}^{m} W_i}{\widehat{\sigma}_W}, \quad \widehat{\sigma}_W^2 = \frac{1}{m-1}\sum_{i=1}^{m}\left(W_i - \frac{1}{m}\sum_{i'=1}^{m} W_{i'}\right)^2.$$

which has an asymptotic t-distribution under H_0 as $m \to \infty$.

Statistical Inference for the Generalization Error

Test II: permutation test

Let $\pi = (\pi_i)_{i=1}^m$ be an element drawn uniformly at random from $\mathcal{B}^m = \{-1, 1\}^m$.

Then, under H_0,

$$T(\pi) = \sum_{i=1}^m W_i \stackrel{\mathscr{D}}{=} \sum_{i=1}^m \pi_i W_i, \quad W_i = I(\widehat{Y}_i \neq Y_i') - I(\widetilde{Y}_i \neq Y_i'),\ i = 1, \ldots, m,$$

where $\stackrel{\mathscr{D}}{=}$ means equality in distribution.

With unlimited computational power, we could generate all 2^m permutations and thereby exactly obtain the (discrete) probability distribution of the random variable $T(\pi)$.

Statistical Inference for the Generalization Error

We could then a define a p-value by

$$1 - \frac{1}{2^m} \sum_{\pi \in \mathcal{B}^m} I\left(\left| \sum_{i=1}^{m} W_i \right| > |T(\pi)| \right)$$

and reject H_0 respectively do not reject H_0 accordingly.

With limited computational power, we sample $\{\pi_{(i)}\}_{i=1}^{N}$ independently and uniformly at random from $\mathcal{B}^m$.

This yields the approximation

$$1 - \frac{1}{N} \sum_{i=1}^{N} I\left(\left| \sum_{i=1}^{m} W_i \right| > |T(\pi_{(i)})| \right)$$

Model Selection

Model selection typically comprises the following aspects:

- deciding on the variables/features to be included,
- choosing the regularization parameter,
- choosing other parameters besides the regularization parameter.

We have not seen an example for th 3rd bullet yet, but we will see several examples later, e.g., when discussing trees and non-linear SVMs.

Model Selection

Model section when given plenty of data:

Data set is split into three subsets:

- one part is used for training,
- the second part (validation set) is used for model selection,
- the third part (test set) is used for model evaluation.

Model Selection

Note that a split into three parts is necessary:

- Doing model selection using the training set will lead to a preference of complex models (overfitting).
- Using the validation set also to estimate the generalization error tends to result into biased estimates.

Model Selection

Using Hoeffding's inequality again, one can show that performing model selection based on a validation set is consistent, i.e., the model with minimum expected risk is chosen with high probability as the size of the validation set grows.

Let $\{f_\mu\}_{\mu \in \mathcal{M}}$ denote the models (binary classifiers) under consideration. Each model index μ may refer to different variables, regularization parameters, etc.

$\widehat{R}_m(f_\mu)$: misclassification error of f_μ on a validation set of size m.

$R(f_\mu)$: expected misclassification error of f_μ, $\mu \in \mathcal{M}$.

Model Selection

Let $f_{\widehat{\mu}}$ denote the model achieving smallest validation error. Then, with probability at least $1 - \delta$, we have

$$R(f_{\widehat{\mu}}) \leq \min_{\mu \in \mathcal{M}} R(f_\mu) + \sqrt{\frac{2}{m} \log\left(\frac{2|\mathcal{M}|}{\delta}\right)}.$$

In other words, as m grows, the expected error of the selected model attains that of the optimal model among those in $\mathcal{M}$.

Model Selection

Proof. (Part I)

We first show that
$|R(f_{\widehat{\mu}}) - \min_{\mu \in \mathcal{M}} R(f_\mu)| \leq 2 \max_{\mu \in \mathcal{M}} |\widehat{R}_m(f_\mu) - R(f_\mu)|$.

Indeed, we have that

$$\begin{aligned} R(f_{\widehat{\mu}}) - \min_{\mu \in \mathcal{M}} R(f_\mu) &= R(f_{\widehat{\mu}}) - \widehat{R}_m(f_{\widehat{\mu}}) + \widehat{R}_m(f_{\widehat{\mu}}) - \min_{\mu \in \mathcal{M}} R(f_\mu) \\ &\leq R(f_{\widehat{\mu}}) - \widehat{R}_m(f_{\widehat{\mu}}) + \max_{\mu \in \mathcal{M}} (\widehat{R}_m(f_\mu) - R(f_\mu)) \\ &\leq 2 \max_{\mu \in \mathcal{M}} |\widehat{R}_m(f_\mu) - R(f_\mu)|. \end{aligned}$$

Note that for the 1st inequality, we use that $f_{\widehat{\mu}}$ minimizes the validation error.

Proof. (Part II)

On the other hand, combining Hoeffding's inequality and Boole's inequality, we obtain

$$\begin{aligned}
&\mathbf{P}\left(\max_{\mu\in\mathcal{M}}|\widehat{R}_m(f_\mu)-R(f_\mu)|>\varepsilon\right)\\
&=\mathbf{P}\left(\bigcup_{\mu\in\mathcal{M}}\{|\widehat{R}_m(f_\mu)-R(f_\mu)|>\varepsilon\}\right)\\
&\leq\sum_{\mu\in\mathcal{M}}\mathbf{P}\left(|\widehat{R}_m(f_\mu)-R(f_\mu)|>\varepsilon\right)\qquad\text{(Boole)}\\
&\leq 2|\mathcal{M}|\exp(-2m\varepsilon^2)\qquad\text{(Hoeffding).}
\end{aligned}$$

The result follows by setting $\varepsilon=\sqrt{\frac{1}{2m}\log(2|\mathcal{M}|/\delta)}$.

Model Selection

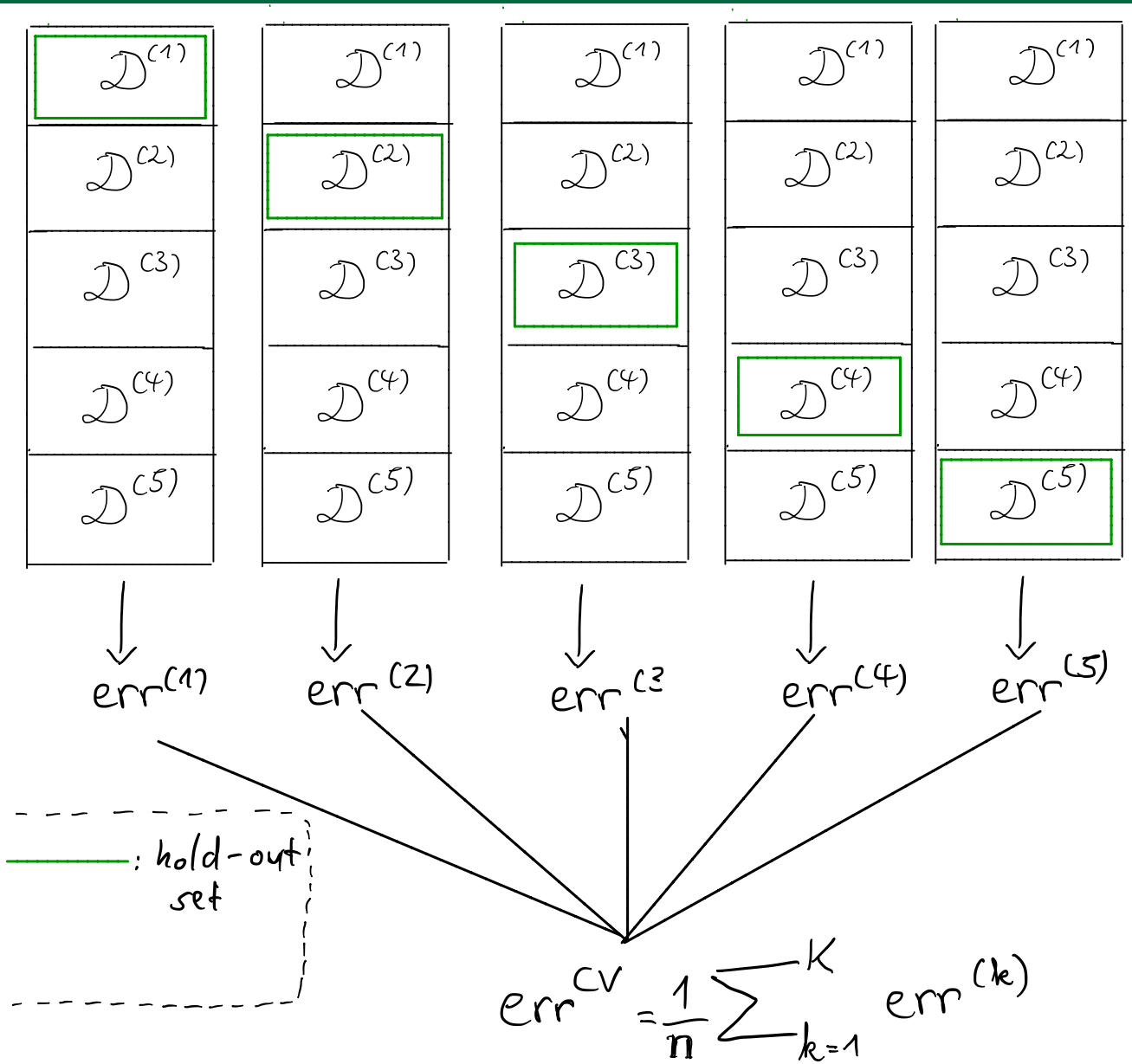
When using a validation set, we may lose a significant number of samples for training.

Cross-validation is one way of avoiding this issue.

The idea of K-fold cross-validation is to partition the training set into K chunks (*folds*).

Training is performed K times so that each fold is left out precisely once as validation set, and the remaining folds are used as training set.

Model Selection



Model Selection

Formal description of K-fold cross-validation:

(1):

Given $\mathcal{D}_n = \{(X_i, Y_i)\}_{i=1}^n$, we partition $\mathcal{D}_n$ into K folds $\mathcal{D}^{(k)} = \{(X_i, Y_i)\}_{i \in \mathcal{I}_k}$, $k = 1, \ldots, K$, of roughly the same size.

- The $\mathcal{I}_k$, $k = 1, \ldots, K$, are disjoint index subsets of $\{1, \ldots, n\}$.

  They are typically generated randomly from a random permutation of $\{1, \ldots, n\}$.

- The number of folds K can be as large n. In this case, one speaks of **Leave-One-Out Cross-Validation**. Most of the time, this is not computationally feasible.

Model Selection

(2):

In the k-th fold, we use $\mathcal{D}^{(k)}$ as test set and the rest $\bigcup_{l \neq k} \mathcal{D}^{(l)}$ as training set based on which we fit a model $\widehat{f}^{(k)}$.

Model $\widehat{f}^{(k)}$ is evaluated on $\mathcal{D}^{(k)}$:

$$\text{err}^{(k)} = \sum_{i \in \mathcal{I}_k} L(Y_i, \widehat{f}^{(k)}(X_i)),$$

where L is some loss function for the problem at hand. For classification, one can use the $0-1$ loss.

Model Selection

(3):
We aggregate the errors

$$\text{err}^{\text{CV}} = \frac{1}{n} \sum_{k=1}^{K} \text{err}^{(k)}.$$

Note: the model that we fit does not change over folds.

To evaluate and compare different models, we need to do all of the above steps for each model under consideration.

Note that pre-processing steps done before model fitting, such as scaling/centering, variable selection etc., need to be done separately in each fold.

Model Selection

(4):
We choose the model with minimum CV error, and then re-fit it with the entire dataset.
Cross-validation can be applied in a variety of situations (regression, classification, density estimation etc.).

Drawbacks:

- Computationally expensive:

  K times more expensive than sample splitting.

  Parallelization requires having copies of the data set on each machine.
- We select a model that does best when using only a fraction of $(K-1)/K$ of the data. For the *entire* training set, this model need not be optimal.

Model Selection

Model Selection in least squares regression without cross-validation: Mallow's C_p.

Let us recall the framework of "fixed X" linear regression with training set

$$Y_i = f^*(x_i) + \sigma\varepsilon_i, \quad i =, \ldots, n,$$

where $\mathbf{E}[\varepsilon_i] = 0$, $\mathbf{E}[\varepsilon_i^2] = \sigma^2$, $i = 1, \ldots, n$, $\{\varepsilon_i\}_{i=1}^n$ uncorrelated.

We perform a least squares fit using a feature expansion in terms of functions $\{\phi_j\}_{j=1}^D$. Given $w \in \mathbb{R}^D$, its MSE is given by

$$\mathsf{MSE}(w) = \mathbf{E}_{\varepsilon'}\left[\frac{1}{n}\sum_{i=1}^n (Y_i' - \Phi(x_i)^\top w)^2\right],$$

where

$$Y_i' = f^*(x_i) + \sigma\varepsilon_i', \quad \varepsilon_i \sim \varepsilon_i', \quad i =, \ldots, n.$$

Model Selection

For an estimator $\widehat{w}_\mu$ obtained from $\{(x_i, Y_i)\}_{i=1}^n$, where $\mu \in \mathcal{M}$ represents some model, we define

$$\mathrm{err}_\mu = \mathbf{E}_\varepsilon \, \mathrm{MSE}(\widehat{w}_\mu) = \mathbf{E}_\varepsilon \, \mathbf{E}_{\varepsilon'} \left[\frac{1}{n} \sum_{i=1}^n (Y_i' - \Phi(x_i)^\top \widehat{w}_\mu)^2 \right].$$

Our goal is to pick the model μ for which err_μ is minimal.

Since err_μ is not known, we need to estimate it.

Model Selection

A natural estimator of err_μ is the training error

$$\widehat{\mathsf{err}}_\mu = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \Phi(x_i)^\top \widehat{w}_\mu)^2.$$

However, this estimator tends to have a significant negative bias, i.e.

$$\mathsf{bias}(\widehat{\mathsf{err}}_\mu) := \mathbf{E}[\widehat{\mathsf{err}}_\mu] - \mathsf{err}_\mu < 0.$$

We can try to estimate the bias and then do a bias correction:

$$\widehat{\widehat{\mathsf{err}}}_\mu = \widehat{\mathsf{err}}_\mu - \widehat{\mathsf{bias}}(\widehat{\mathsf{err}}_\mu),$$

where $\widehat{\mathsf{bias}}(\widehat{\mathsf{err}}_\mu)$ is a suitable estimator of the bias.

Model Selection

It can be shown that for any estimator satisfying

$$\mathbf{\Phi}\widehat{w}_\mu = A_\mu \mathbf{y},$$

for a linear map $A_\mu : \mathbb{R}^n \to \mathbb{R}^n$, we have

$$\mathsf{bias}(\widehat{\mathsf{err}}_\mu) = -2\sigma^2 \operatorname{tr}(A_\mu)/n.$$

This yields Mallow's C_p criterion for model selection:

$$C_p(\mu) = \underbrace{\frac{1}{n}\sum_{i=1}^{n}(Y_i - \Phi(x_i)^\top \widehat{w}_\mu)^2}_{\widehat{\mathsf{err}}_\mu} + 2\sigma^2 \frac{\operatorname{tr}(A_\mu)}{n}.$$

Model Selection

$$C_p(\mu) = \underbrace{\frac{1}{n}\sum_{i=1}^{n}(Y_i - \Phi(x_i)^\top \widehat{w}_\mu)^2}_{\widehat{\mathsf{err}}_\mu} + 2\sigma^2 \frac{\mathrm{tr}(A_\mu)}{n}.$$

For least squares fits with each model μ representing a different subset of variables, we have

$$A_\mu = \mathbf{\Phi}_\mu(\mathbf{\Phi}_\mu^\top \mathbf{\Phi}_\mu)^{-1}\mathbf{\Phi}_\mu^\top,$$
$$\Rightarrow \mathrm{tr}(A_\mu) = \text{number of columns (variables) in } \mathbf{\Phi}_\mu.$$

This makes sense: Mallow's C_p penalizes more complex models and encourages parsimonious models.

Model Selection

$$C_p(\mu) = \underbrace{\frac{1}{n}\sum_{i=1}^{n}(Y_i - \Phi(x_i)^\top \widehat{w}_\mu)^2}_{\widehat{\mathsf{err}}_\mu} + 2\sigma^2 \frac{\mathrm{tr}(A_\mu)}{n}.$$

For ridge regression, the set of models $\mathcal{M}$ can be identified with a set of values Λ for the ridge parameter λ. We have

$$A_\mu \equiv A_\lambda := \mathbf{\Phi}(\mathbf{\Phi}^\top \mathbf{\Phi} + n\lambda I)^{-1}\mathbf{\Phi}^\top$$

$$\Rightarrow \mathrm{tr}(A_\lambda) = \sum_{j=1}^{D} \frac{\lambda_j}{\lambda_j + \lambda n},$$

where $\{\lambda_j\}_{j=1}^D$ are the eigenvalues (= squared singular values) of $\mathbf{\Phi}^\top \mathbf{\Phi}$.

Model Selection

$$C_p(\mu) = \underbrace{\frac{1}{n}\sum_{i=1}^{n}(Y_i - \Phi(x_i)^\top \widehat{w}_\mu)^2}_{\widehat{\mathsf{err}}_\mu} + 2\sigma^2 \frac{\mathrm{tr}(A_\mu)}{n}.$$

A serious issue about the C_p is that σ^2 is not known, hence needs to be estimated. Unfortunately, this is far from straightforward.

The "classical" estimator of σ^2 is given by

$$\widehat{\sigma}^2 = \frac{1}{n-D}\sum_{i=1}^{n}(Y_i - \Phi(x_i)^\top \widehat{w}_{\mathsf{LS}})^2,$$

where $\widehat{w}_{\mathsf{LS}}$ is the least squares estimator of the largest model (D variables) under consideration.

Model Selection

The estimator

$$\widehat{\sigma}^2 = \frac{1}{n-D}\sum_{i=1}^{n}(Y_i - \Phi(x_i)^\top \widehat{w}_{\mathsf{LS}})^2,$$

is consistent only under restrictive conditions:

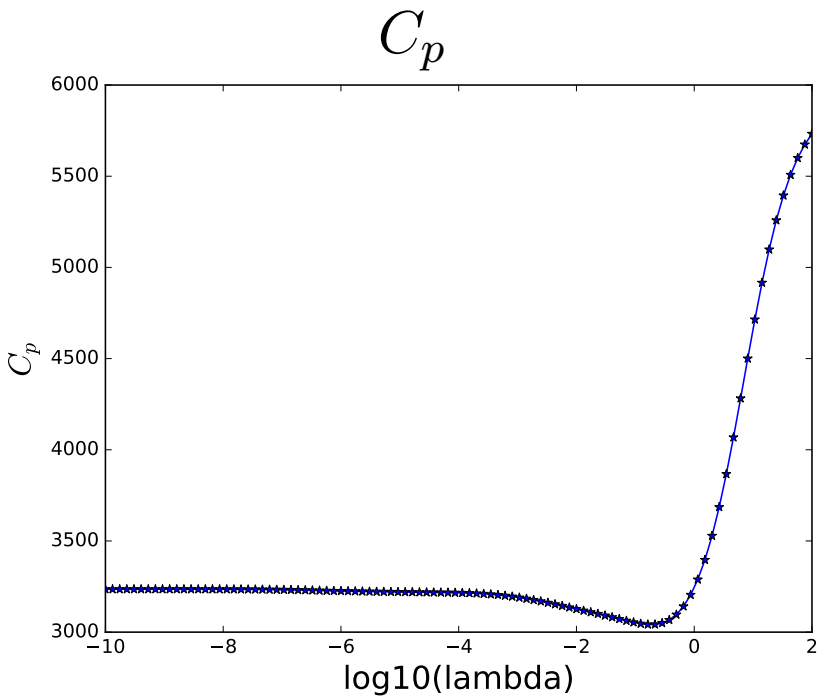
- the linear model holds exactly (i.e., $\boldsymbol{f}^* = \boldsymbol{\Phi} w^*$ for some $w^* \in \mathbb{R}^D$),
- n is large relative to D.

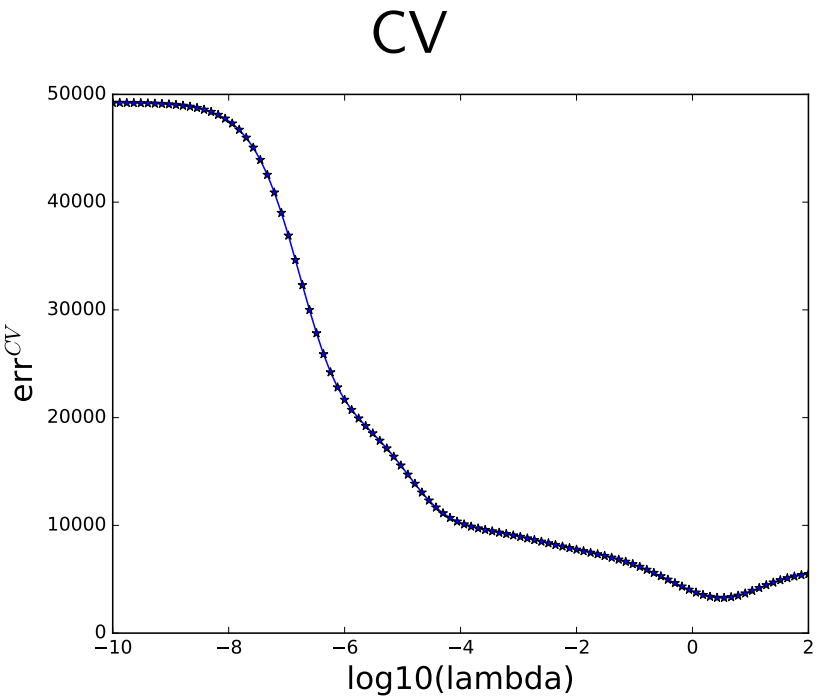
Model Selection

Example I: choosing the ridge parameter for the Pima Indian Diabetes data set ($n = 442$, $D = 64$).

1) via five-fold cross-validation,
2) via Mallow's C_p, with σ^2 replaced by the estimator discussed on the previous slides.

Model Selection



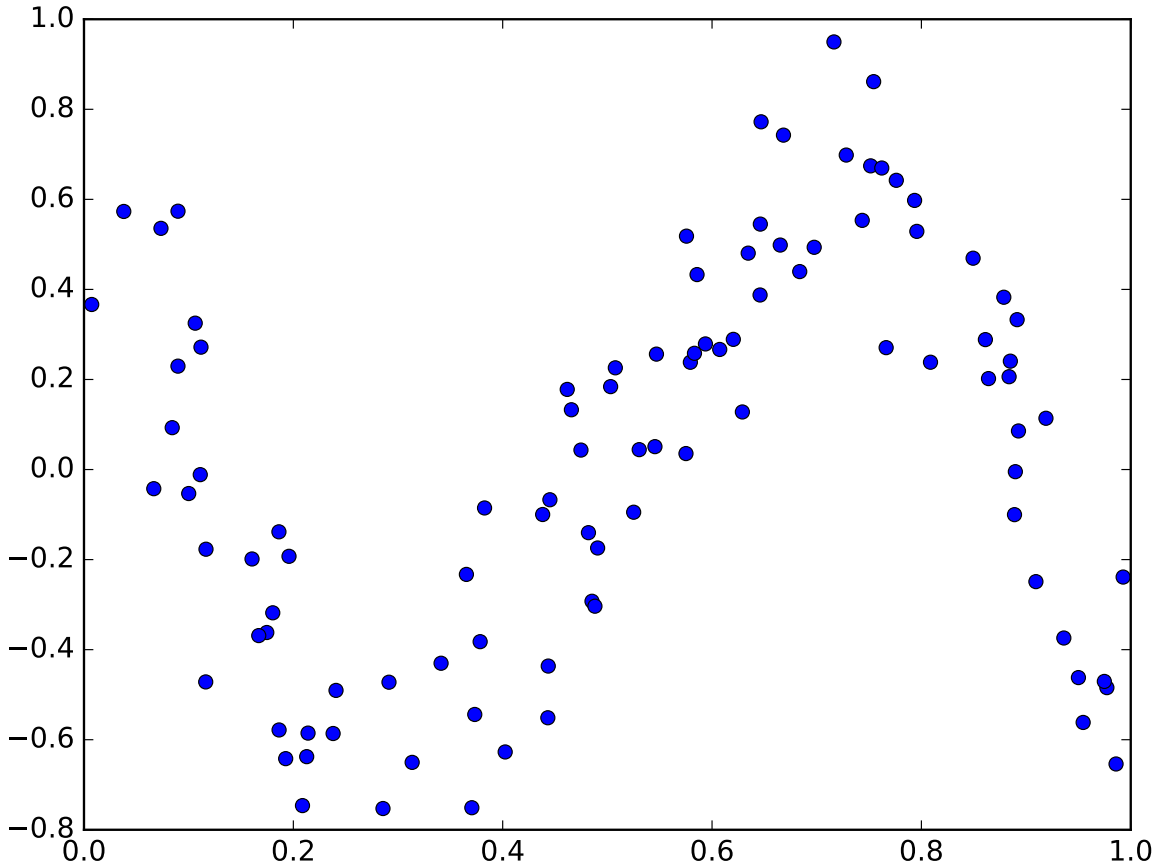


Both curves reach a dip, but at different locations:

$$\lambda_{C_p} = 0.16, \qquad \lambda_{\text{CV}} = 3.51.$$

Model Selection

Example II: choosing the regularization parameter in nonparametric regression with a penalty encouraging smoothness.



Model Selection

Our approach was to develop the unknown function f^* in a (truncated) Fourier series with maximum frequency N:

$$\min_{f \in \mathcal{F}_N} \sum_{i=1}^{n} \{Y_i - f(X_i)\}^2 + \lambda \Omega(f), \quad \Omega(f) = \int_0^1 \{f'(x)\}^2 \, dx.$$

This was shown to be equivalent to a least squares problem with weighted ridge regularizer with weights $\{\omega_j\}_{j=1}^D$

$$\min_{\alpha \in \mathbb{R}^D} \|\mathbf{y} - \mathbf{\Phi}\alpha\|_2^2 + \lambda \sum_{j=1}^{D} \omega_j \alpha_j^2,$$

and $D = 2N + 1$.

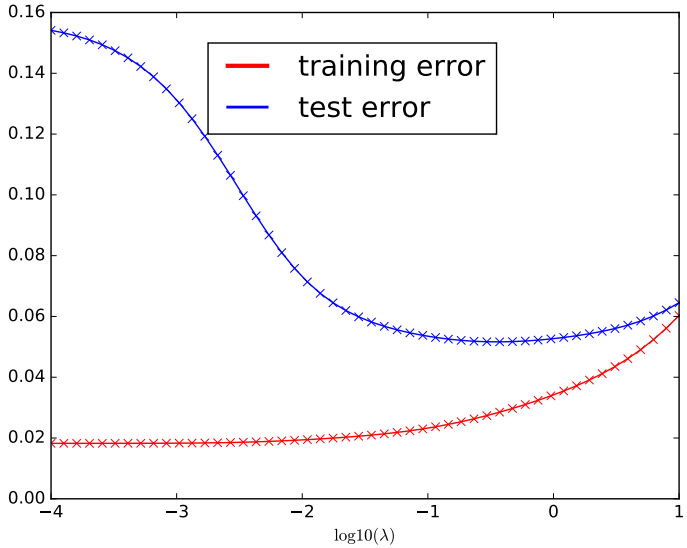
Model Selection

Let $\widehat{\alpha}$ denote the optimal set of coefficients for the problem on the previous slide. We have

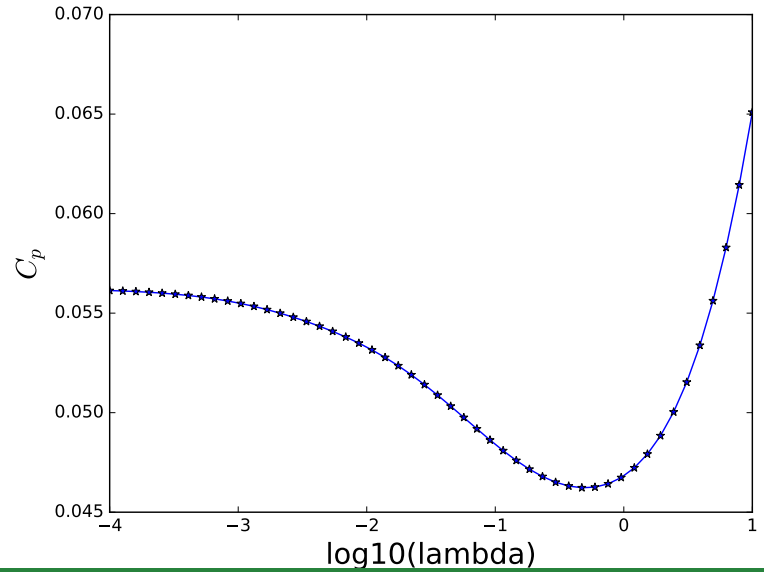
$$\begin{aligned}\boldsymbol{\Phi}\widehat{\alpha} &= \underbrace{\boldsymbol{\Phi}(\boldsymbol{\Phi}^\top\boldsymbol{\Phi} + \lambda\boldsymbol{\Omega})^{-1}\boldsymbol{\Phi}^\top}_{A_\lambda}\mathbf{y}, \quad \boldsymbol{\Omega} = \mathsf{diag}(\omega_1, \ldots, \omega_D) \\ &= A_\lambda \mathbf{y},\end{aligned}$$

Hence, the C_p remains applicable. We fix $N = 25$ and estimate σ^2 from the solution with $\lambda = 0$.

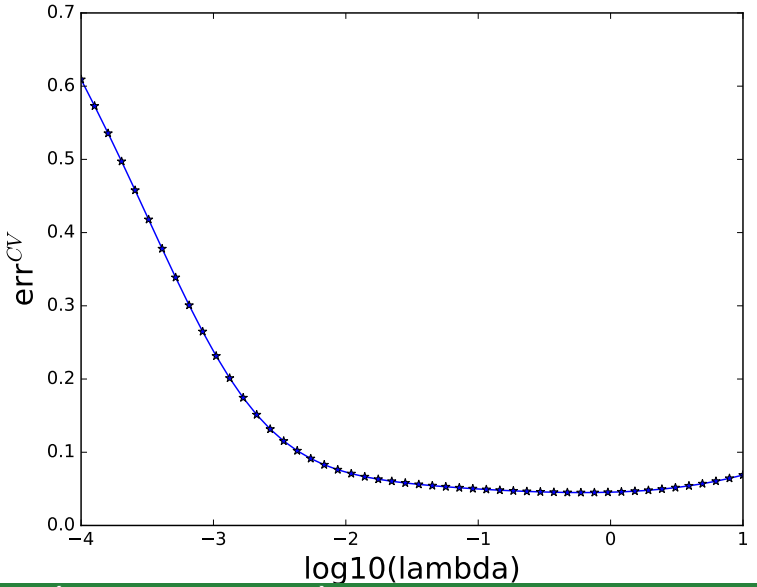
Model Selection



C_p



CV



Model Selection

In this example, both the C_p and five-fold CV work rather well:

- the λ achieving optimal test error of 0.0516 equals 0.37,
- the C_p-optimal λ equals 0.47 with a test error of 0.0517,
- the CV-optimal λ equals 0.6 with a test error of 0.0519.

Model Selection

Mallow's C_p can be linked to Akaike's information criterion (AIC).

The AIC is a general approach for doing model selection for parametric models, i.e., it is assumed that the data $\mathcal{D}_n$ are generated from some parametric family $\{P_\theta,\ \theta \in \Theta\}$.

Example I:

$Y_i|\Phi(X_i) \sim N(\Phi(X_i)^\top w, \sigma^2)$, $i = 1, \ldots, n$,

$\{Y_i\}_{i=1}^n$ (conditionally) independent.

Parameter: w. Inference by maximum likelihood is equivalent to least squares estimation.

Model Selection

Example II:

$Y_i|\Phi(X_i) \sim$ Bernoulli $\left(\frac{\exp(\Phi(X_i)^\top w)}{1+\exp(\Phi(x_i)^\top w)}\right)$, $i = 1, \ldots, n$,

$\{Y_i\}_{i=1}^n$ (conditionally) independent.

Parameter: w. Inference by maximum likelihood (logistic regression) is equivalent to to using the logistic loss in binary classification.

Model Selection

The AIC is defined by

$$\max_{\theta \in \Theta_\mu} \{-2\text{log-likelihood}(\theta)\} + 2\dim(\Theta_\mu),$$

where $\Theta_\mu \subseteq \Theta$ is a subset (subspace) of the parameter set corresponding to some model $\mu \in \mathcal{M}$.

$\dim$: means "dimension" respectively the $\#$ of variables.

The AIC naturally applies to variable selection:

- Θ corresponds to the "full model"
- $\{\Theta_\mu\}_{\mu \in \mathcal{M}}$ correspond to sub-models, with certain variables left out.

Model Selection

For Example I (Gaussian regression), it turns out that model selection based on the AIC is equivalent to using Mallow's C_p.

A third criterion which is closely related is the Bayesian Information Criterion (BIC):

$$\max_{\theta \in \Theta_\mu} \{-2\text{log-likelihood}(\theta)\} + 2 \cdot \log(n) \cdot \dim(\Theta_\mu),$$

The BIC has better asymptotic properties with regard to consistency of model selection.

Model Selection

There are many more model selection criteria in the literature:

- GCV (generalized cross-validation),
- AIC_c (corrected AIC),
- RIC (Risk Inflation criterion),
- Stein's Unbiased Risk Estimate (SURE),
- $\vdots$